正交投影

正交投影 3

对任何给定的向量 $\mathbf{v} \in V$, 其满足约束 $\mathbf{u} \in U \subset V$ 的最佳逼近为 $\mathbf{v}$ 在 U 上的投影 $P_U\mathbf{v}$:

$$P_U\mathbf{v} = arg\min_{\mathbf{u}\in U} \|\mathbf{v}-\mathbf{u}\|^2$$

3.1 背景 1
笛卡尔坐标 1
线性方程组 2
3.2 向量空间 3
向量空间的定义 3
函数空间 4
有限维向量空间 5
3.3 内积向量空间 6
正交投影 8
3.4 随机变量空间 11
3.5 欧氏空间的投影矩阵 15
奇异值分解 15
广义逆 17
正交投影矩阵 18

初等代数对实数进行四则运算，高等代数则将代数运算的对象拓展到一般集合的元素。集合中赋予的一种或多种运算规则称为代数结构，定义了某种代数结构并且关于运算封闭的集合称为空间。定义了集合元素之间的 “加法” 和实数（或复数）与集合元素的 “乘积”，且满足通常的线性运算规则的封闭集合称为向量空间或线性空间，集合元素称为向量。特别地，线性代数研究有限维向量空间中的线性运算、以及空间之间的线性变换及其矩阵表示，泛函分析和概率论研究实函数向量空间上的线性运算和分析性质。

3.1-3.3节回顾向量空间和内积向量空间的投影，熟悉相关内容的读者可以略过。3.4节讨论随机变量空间，以投影观点看待条件期望以及总体线性回归模型；3.5 节介绍欧氏空间中的投影，并以此观点看待样本线性回归模型的最小二乘法。

3.1 背景

3.1.1 笛卡尔坐标

向量的概念最初来自于平面或立体笛卡尔坐标系中的坐标表示，也与线性方程组的研究有关。笛卡尔将实数的四则运算拓展到二维 R^2 空间或者三维 R^3 空间的坐标向量。以 R^2 空间为例，平面上任何一点可由两个实数确定，用二维笛卡尔坐标表示为向量 $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, a_1, a_2 \in R$。实数标量 $\lambda \in R$ 与向量 $\mathbf{a}$ 的乘积（即数乘）以及向量 $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ 与向量 $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ 之和分别定义为

$$\lambda\mathbf{a} = \begin{pmatrix} \lambda a_1 \\ \lambda a_2 \end{pmatrix}, \quad \mathbf{a}+\mathbf{b} = \begin{pmatrix} a_1+b_1 \\ a_2+b_2 \end{pmatrix}$$

图 3.1: 勒内 · 笛卡尔 René Descartes (1596-1650)，法国数学家、哲学家。1637 年笛卡尔创立了直角坐标系，为微积分的建立奠定了基础，并形成了以代数方法研究几何的解析几何或坐标几何方法。

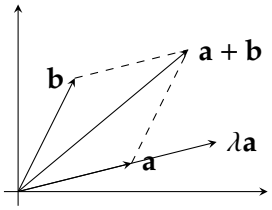


图 3.2: R^2 中向量的加法和数乘构成了平面上点（或向量）的代数运算。

图3.2是向量数乘和向量加法的演示。

若进一步定义内积、模长，则我们可以利用这些代数工具研究平面或空间上的几何，从而数与形达到了完美统一。进而向量的概念被推广到 R^n 甚至一般的形式。长度为 n 的实数向量的全体构成集合 R^n：

$$R^n = \{\mathbf{x} = (x_1, ..., x_n)^\top : x_1, ..., x_n \in R\}$$

类似地定义加法和数乘之后，R^n 就构成了最常见的向量空间。线性代数将 R^n 拓展到一般数学对象并定义类似的运算法则。虽然习惯上向量通常指的是 R^n 中 n 个实数堆积成的数学对象 $\mathbf{x} = (x_1, ..., x_n)^\top \in R^n$。但当讨论向量空间时，向量可以是任何数学对象。

3.1.2 线性方程组

线性代数中一些概念的定义特别是四个基本空间（行、列的核空间、像空间）以及相关概念来自于线性方程组的求解问题。假设有 m 元一次线性方程组（n 个方程）

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = b_1 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = b_n \end{cases},$$

横向看每个方程，它们构成 m 维空间的 n 个超平面，方程的解为所有超平面的交点。记

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \mathbf{a}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}, j = 1, ..., m$$

则线性方程组表示为

$$\mathbf{a}x_1 + \mathbf{a}_2x_2 + ... + \mathbf{a}_mx_m = \mathbf{b},$$

我们称 $\mathbf{b}$ 是 $\mathbf{a}_1, ..., \mathbf{a}_m$ 的线性组合。方程有解当且仅当 $\mathbf{b}$ 可由 $\mathbf{a}_1, ..., \mathbf{a}_m$ 线性组合而成，即 $\mathbf{b}$ 属于 $\mathbf{a}_1, ..., \mathbf{a}_m$ 张成的向量空间

$$C(A) = \{\mathbf{a}x_1 + \mathbf{a}_2x_2 + ... + \mathbf{a}_mx_m : x_1, ..., x_m \in R\},$$

若 $\mathbf{a}_1, ..., \mathbf{a}_m$ 线性无关，即 $\mathbf{a}x_1 + \mathbf{a}_2x_2 + ... + \mathbf{a}_mx_m = 0$ 一定蕴含 $x_1 = ... = x_m = 0$, 则 $\mathbf{a}_1, ..., \mathbf{a}_m$ 称为 $C(A)$ 的一组基，而 $\mathbf{x}$ 是 $\mathbf{b}$ 在该组基下的坐标。这导致人们研究一般意义上的向量张成的向量子空间。

进一步，引入矩阵即矩阵乘法，记矩阵 A, 向量 $\mathbf{x}$ 分别为

$$A = (\mathbf{a}_1, ..., \mathbf{a}_m) = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix},$$

则线性方程组简写为

$$A\mathbf{x} = \mathbf{b},$$

我们称 $\mathbf{b}$ 是 $\mathbf{x}$ 的线性变换。该方程，$\mathbf{x}$ 是 A 的列为基的空间中的坐标。方程何时有唯一解？若方程有两个解 $\mathbf{x}_1, \mathbf{x}_2$, 即 $A\mathbf{x}_1 = \mathbf{b}$, $A\mathbf{x}_2 = \mathbf{b}$, 则 $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$, 这引申出考察核空间 $N(A) = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$ 除了包含 $\mathbf{0}$ 之外是否还有其它元素。

3.2 向量空间

向量和向量空间是线性代数的基本结构框架。向量空间也称为线性空间，是对线性运算封闭的数学对象的集合，集合中的元素或数学对象称为向量。

3.2.1 向量空间的定义

类似于笛卡尔坐标系中向量的代数运算，一般向量空间及其代数运算定义如下。

定义 3.1 如果一个集合 V 配备如下两个二元运算（加法和数乘）并满足如下的运算法则，则称 V 是一个向量空间或线性空间，V 的元素称为向量。

二元运算

- 加法：任何 $\mathbf{u}, \mathbf{v} \in V$ 对应于唯一一个 V 的元素，记作 $\mathbf{u} + \mathbf{v} \in V$.
- 数乘：任何 $\mathbf{v} \in V$ 和任何实数 $\lambda \in R$（也可以是任何其它数域，比如复数域）对应于 V 中唯一的一个元素，记作 $\lambda\mathbf{v} \in V$.

运算法则：对任何 $\mathbf{u},\mathbf{v},\mathbf{w}\in V$ 和任何实数 $\lambda,\gamma\in R$，前述二元运算满足如下法则：

- 加法交换律 $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$.
- 加法结合律：$(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$.
- 存在加法单位元(记作 $\mathbf{0}$): 存在元素 $\mathbf{0}\in V$, 使得 $\mathbf{0}+\mathbf{v}=\mathbf{v}$.
- 存在加法逆：对任何 $\mathbf{v}\in V$, 存在 $\mathbf{u}\in V$ 使得 $\mathbf{v}+\mathbf{u}=\mathbf{0}$, 记作 $\mathbf{u}=-\mathbf{v}$.
- 数乘分配律: $\lambda(\mathbf{u}+\mathbf{v})=\lambda\mathbf{u}+\lambda\mathbf{v}$, $(\lambda+\gamma)\mathbf{v}=\lambda\mathbf{v}+\gamma\mathbf{v}$.
- 数乘结合律: $(\lambda\gamma)\mathbf{v}=\lambda(\gamma\mathbf{v})$.
- 数字 $1\in R$ 是数乘单位元：$1\mathbf{v}=\mathbf{v}$.

显然，3.1节中笛卡尔坐标空间 R^n 中的加法和数乘满足上述定义中的要求，因而 R^n 构成一个向量空间。除了 R^n 之外，最常见的向量空间是函数空间（实际上，R^n 也是一个函数空间, 参见例 3.2)。

定义 3.2 若 V,W 是两个向量空间，映射 $T:V\to W$ 称为线性映射或线性变换，如果对任何 $\lambda_1,\lambda_2\in R$ 和任何 $\mathbf{u},\mathbf{v}\in V$ 有

$$T(\lambda_1\mathbf{u}+\lambda_2\mathbf{v})=\lambda_1T\mathbf{u}+\lambda_2T\mathbf{v}.$$

所有从 V 到 W 的线性映射记作 $\mathcal{L}(V,W)$。特别地，$\mathcal{L}(V,R)$ 中的线性映射称为线性泛函（functional），$\mathcal{L}(V,V)$ 中的线性映射称为线性算子。

例 3.1 对任何 $\mathbf{x}=(x_i),\mathbf{y}=(y_i)\in R^n$，向量加法和数乘分别定义为

$$\mathbf{x}+\mathbf{y}=(x_i+y_i),\lambda\mathbf{x}=(\lambda x_i),\lambda\in R.$$

则 R^n 是一个向量空间。赋予内积的向量空间 R^n 称为欧氏空间。

3.2.2 函数空间

某个集合上的所有实函数在定义了函数加法和数乘之后构成一个向量空间, 称为函数空间，是泛函分析和概率论研究的主要框架结构。下面的定义和例子来自于 Axler 所著线性代数教材[1]，他引入记号 R^S 代表集合 S 上的实函数向量空间。

定义 3.3 (函数空间 R^S) 假设 S 是任一非空集合，记所有 S 到实数域 R 的函数集合

$$R^S=\{f:S\to R\}.$$

定义加法和数乘运算如下

▶ 对任何 $f, g \in R^S$, 定义 $f+g$ 为函数

$$(f+g)(x) = f(x) + g(x), \forall x \in S.$$

▶ 对任何 $f \in R^S$ 和 $\lambda \in R$, 定义数乘 λf 为函数

$$(\lambda f)(x) = \lambda f(x), \forall x \in S.$$

容易验证上述运算满足定义3.1 中的法则。另外，0 映射 ($f(x) \equiv 0$) 是加法单位元，$-f$ 是 $f \in R^S$ 的加法逆。所以 R^S 是一个向量空间。

例 3.2 下面是常见的函数空间。

(1) $S = R$ 时, R^R 为实轴上所有实函数的集合。当函数性质具有某些特殊的限制时，比如平方可积函数构成一个 R^R 的向量子空间，通常记作 L^2 或 $L^2(R)$；

(2) $S = \{1, ..., n\}$, 则 $R^S = R^{\{1,...,n\}} = \{x : S \to R\}$ 是 S 上的全体实函数。任一函数 $x \in R^{\{1,...,n\}}$ 以其所有 n 个取值 $(x(1), ..., x(n))$ 或 $(x_1, ..., x_n)$ 表示（习惯上，当定义域有限时，以 x_i 而不是 $x(i)$ 表示函数值），所以 R^n 中的向量 $(x_1, ..., x_n)^\top$ 可理解为 $S = \{1, ..., n\}$ 上的函数 x，即

$$R^{\{1,...,n\}} = \{x : \{1, .., n\} \to R\} = \{(x_1, ..., x_n)^\top : x_i \in R\} = R^n.$$

同样地，无穷序列 $(x_1, x_2, ...)$ 可理解成 $\{1, 2, ..., \}$ 上的函数 $x : \{1, 2, ..., \} \to R$, 而 $R^{\{1,2,...,\}}$ 即是通常的 R^∞。

(3) $S = \Omega$ (随机结果集合/样本空间), R^Ω 为随机结果到实数域 R 的函数。因为对随机结果集合引入了概率测度且不是所有 Ω 的子集都能计算概率，这导致需要对 R^Ω 中函数的复杂性施加某种 (可测，可度量) 限制，称为随机变量（参见3.4）。一个随机变量 x 是 Ω 到 R 的函数 $x : \omega \in \Omega \to R$，它落在任一区间[1]的随机结果集合 $\{\omega \in \Omega : x(\omega) \in (s, t], s \le t \in R\} \in \mathcal{F}$, 其中 $\mathcal{F}$ 是所有可以计算概率的随机结果的集合，称为 σ 域。我们记所有平方可积的随机变量组成的空间为 $L^2(\mathcal{F}) \subset R^\Omega$。

1: 为什么考虑区间的原像? 区间, 特别是无穷小区间是研究分析性质的工具

3.2.3 有限维向量空间

给定有限个向量 $\mathbf{v}_1, ..., v_m \in V$, 它们张成的线性子空间是所有可能的线性组合

$$C(\mathbf{v}_1, ..., \mathbf{v}_m) = \{\lambda_1 \mathbf{v}_1 + ... + \lambda_m \mathbf{v}_m : \lambda_1, ..., \lambda_m \in R\} \subset V$$

如果线性（子）空间可由 n 个线性无关的向量张成，则该空间是 n 维线性空间，这 n 个向量是该空间的基。线性空间中的任一向量可由该空间的基唯一确定，基的组合系数称为坐标。

假设两个有限维向量空间 V, W 的维数分别是 m, n, 它们的基分别是 $\{\mathbf{v}_1, ..., \mathbf{v}_m\}$ 和 $\{\mathbf{w}_1, ..., \mathbf{w}_n\}$。假设线性变换 $T \in \mathcal{L}(V, W)$, 因为 $T(\mathbf{v}_j) \in W, 1 \leq j \leq m$, 故存在唯一坐标或组合系数 $a_{1j}, ..., a_{nj} \in R$, 使得

$$T(\mathbf{v}_j) = a_{1j}\mathbf{w}_1 + ... + a_{nj}\mathbf{w}_n$$

这 m 组系数按列排列成 $n \times m$ 矩阵

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

形式上记作

$$T(\mathbf{v}_1, ..., \mathbf{v}_m) = (T(\mathbf{v}_1), ..., T(\mathbf{v}_m)) = (\mathbf{w}_1, ..., \mathbf{w}_n) \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

最后一式可看作是行向量域矩阵的乘积。该矩阵唯一确定了变换 T, 是在两组基给定的条件下线性变换的矩阵表示。例如, R^n 是 n 维向量空间, 其标准基为 $\mathbf{e}_1 = (1, 0, ..., 0)^\top, \mathbf{e}_2 = (0, 1, ..., 0)^\top, \mathbf{e}_n = (0, 0, ..., 1)^\top$。实际上, 任何 n 维向量空间都可以看作是 R^n (同构)。

3.3 内积向量空间

类似于笛卡尔坐标系中, 我们可以定义向量空间中任何两个向量之间的内积, 用于度量向量之间的相似性, 并进而考察向量的长度、向量之间的夹角等几何性质。赋予内积的向量空间称为内积向量空间。

定义 3.4 假设 V 是一个向量空间, 映射 $\langle \cdot, \cdot \rangle : V \times V \to R$ 称为是一个内积, 如果满足如下条件:

- 正性: 对任何 $\mathbf{v} \in V, \langle \mathbf{v}, \mathbf{v} \rangle \geq 0$, 且 $\langle \mathbf{v}, \mathbf{v} \rangle = 0 \Leftrightarrow \mathbf{v} = 0$。$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ 称为 $\mathbf{v}$ 的范数或模长。
- 对称性: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$, $\forall \mathbf{u}, \mathbf{v} \in V$.
- 双线性: 对固定的 $\mathbf{w} \in V, \langle \cdot, \mathbf{w} \rangle$ 和 $\langle \mathbf{w}, \cdot \rangle$ 都是 $V \to R$ 的线性泛函。具体地, 对任何 $\lambda_1, \lambda_2 \in R, \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, $\langle \lambda_1 \mathbf{u} + \lambda_2 \mathbf{v}, \mathbf{w} \rangle = \lambda_1 \langle \mathbf{u}, \mathbf{w} \rangle + \lambda_2 \langle \mathbf{v}, \mathbf{w} \rangle$。

定义 3.5 若 $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, 称 $\mathbf{u}, \mathbf{v}$ 正交, 记作 $\mathbf{u} \perp \mathbf{v}$。假设 U 是向量空间 V 的子集, 若向量 $\mathbf{v} \perp \mathbf{u}, \forall \mathbf{u} \in U$, 则称 $\mathbf{v}$ 与 U 正交, 记作 $\mathbf{v} \perp U$。

定理 3.1 (毕达哥拉斯定理，勾股定理) 假设向量空间中 $\mathbf{u} \perp \mathbf{v}$, 则

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

内积可以自然地诱导出距离：任何 $\mathbf{u}, \mathbf{v} \in V$ 的距离定义为：

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|.$$

定义了距离就可以考虑向量空间的极限、积分等分析性质。如果一个空间对极限封闭（完备），称为希尔伯特空间。

定义 3.6 假设 V 是内积向量空间，假设向量序列 $\mathbf{v}_i, i = 1, 2, \ldots \in V$ 是一个 Cauchy 序列，即 $d(\mathbf{v}_n, \mathbf{v}_m) = \|\mathbf{v}_n - \mathbf{v}_m\| \to 0, n, m \to \infty$, 如果存在 $\mathbf{v} \in V$ 使得 $d(\mathbf{u}_n, \mathbf{v}) = \|\mathbf{u}_n - \mathbf{v}\| \to 0$，则称 V 是完备的，V 称为希尔伯特空间。任何有限维内积向量空间是希尔伯特空间。

例 3.3 下面是几个常见的内积向量空间，它们都是希尔伯特空间。

(1) **欧氏空间** R^n.
对任何 $\mathbf{x} = (x_i), \mathbf{y} = (y_i) \in R^n$，定义内积

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i,$$

容易验证它满足内积的要求，而 $\|\mathbf{x}\| = (\mathbf{x}^\top \mathbf{x})^{1/2} = (\sum_{i=1}^n x_i^2)^{1/2}$ 称为 $\mathbf{x}$ 的模长。赋予内积的向量空间 R^n 称为欧氏空间。

(2) **函数空间** L^2.
平方可积函数空间 $L^2 \subset = \{f : R \to R | \inf f^2(x)dx < \infty\}$, 对任何 $f, g \in L^2$, 定义内积

$$\langle f, g \rangle = \int f(x)g(x)dx$$

模长 $\|f\| = (\int f(x)^2 dx)^{1/2}$。

(3) **随机变量空间** $L^2(\mathcal{F})$.
假设概率空间 $(\Omega, \mathcal{F}, P)$, 对任何二阶矩存在的随机变量 (即 $\mathcal{F}$-可测函数) x, y, 定义内积

$$\langle x, y \rangle = E(xy)$$

模长 $\|x\| = (Ex^2)^{1/2}$。

3.3.1 正交投影

定义 3.7 假设 U 是内积向量空间 V 的向量子空间，对任一 $\mathbf{v} \in V$，如果存在 $\hat{\mathbf{v}} \in U$ 使得

$$\mathbf{v} - \hat{\mathbf{v}} \perp U, \text{ 即 } \langle \mathbf{v} - \hat{\mathbf{v}}, \mathbf{u} \rangle = 0, \forall \mathbf{u} \in U.$$

$\hat{\mathbf{v}}$ 称为 $\mathbf{v}$ 在 U 上的正交投影，记为 $\hat{\mathbf{v}} = P_U \mathbf{v}$。

很多二次极小化问题可以描述为：对任何给定的向量 $\mathbf{v} \in V$，求解满足某种约束的向量 $\mathbf{u}$ 用于逼近 $\mathbf{v}$，使得逼近误差最小。如果误差大小以两者的距离平方衡量则（称为最小二乘问题），而约束可以表示成 $\mathbf{u} \in U$，其中 $U \subset V$ 是一个线性子空间，则最优解是投影 $P_U \mathbf{v}$。

定理 3.2 (最小二乘) 任何 $\mathbf{v} \in V$ 在子空间 U 上的正交投影 $\hat{\mathbf{v}} = P_U \mathbf{v}$ 是 U 空间中距离 $\mathbf{v}$ 最近的向量：

$$\hat{\mathbf{v}} = arg \min_{\mathbf{u} \in U} \|\mathbf{u} - \mathbf{v}\|^2$$

证明： 任取 $\mathbf{u} \in U$，因为 $\hat{\mathbf{v}} \in U$，所以 $\mathbf{u} - \hat{\mathbf{v}} \in U$，由 $\mathbf{v} - \hat{\mathbf{v}} \perp U \Rightarrow \mathbf{v} - \hat{\mathbf{v}} \perp \mathbf{u} - \hat{\mathbf{v}}$，所以

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u} - \hat{\mathbf{v}} + \hat{\mathbf{v}} - \mathbf{v}\|^2 = \|\mathbf{u} - \hat{\mathbf{v}}\|^2 + \|\hat{\mathbf{v}} - \mathbf{v}\|^2 \geq \|\hat{\mathbf{v}} - \mathbf{v}\|^2$$

■

下述命题表明若 U 是一维向量子空间（即某个向量张成的空间），则 $P_U \mathbf{v}$ 存在。

命题 3.3 假设 V 是一个内积及向量空间，假设 $\mathbf{v} \in V$ 是任何一个向量。假设 $\mathbf{u} \neq 0 \in V$，子空间 $U = C(\mathbf{u}) = \{\lambda \mathbf{u} : \lambda \in R\} \subset V$。则 $P_{\mathbf{u}} \mathbf{v} = \left(\frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \right) \mathbf{u}$ 是 $\mathbf{v}$ 在 U 上的投影。

证明： 对任何 $\mathbf{v} \in V$，假设存在某个 $\lambda \in R$ 使得 $P_{\mathbf{u}} \mathbf{v} = \lambda \mathbf{u}$，由

$$0 = \langle \mathbf{v} - \lambda \mathbf{u}, \mathbf{u} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle - \lambda \langle \mathbf{u}, \mathbf{u} \rangle$$

所以 $\lambda = \langle \mathbf{v}, \mathbf{u} \rangle / \langle \mathbf{u}, \mathbf{u} \rangle$，从而 $\lambda \mathbf{u} = \left(\frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \right) \mathbf{u}$. 是 $\mathbf{v}$ 在一维向量子空间 $U = C(\mathbf{u})$ 上的投影。 ■

注：当 $V = R^n$ 时，对任何 $\mathbf{u}, \mathbf{v} \in R^n$

$$P_{\mathbf{u}} \mathbf{v} = \left(\frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \right) \mathbf{u} = \left(\frac{\mathbf{u}\mathbf{u}^\top}{\mathbf{u}^\top \mathbf{u}} \right) \mathbf{v}$$

其中 $P_{\mathbf{u}} = \mathbf{u}\mathbf{u}^{\top}/\mathbf{u}^{\top}\mathbf{u} = \mathbf{u}(\mathbf{u}^{\top}\mathbf{u})^{-1}\mathbf{u}^{\top}$ 是向量 $\mathbf{u}$ 对应的投影矩阵。

定理 3.4 (Cauchy-Schwarz 不等式) 对任何 $\mathbf{u}, \mathbf{v} \in V$,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|,$$

等号成立当且仅当 $\mathbf{u} = \lambda \mathbf{v}$（某个 $\lambda \in R$）。

证明：由命题 3.3, $\hat{\mathbf{v}} = P_{\mathbf{u}}\mathbf{v} = (\langle \mathbf{v}, \mathbf{u} \rangle / \langle \mathbf{u}, \mathbf{u} \rangle)\, \mathbf{u}$, 记 $\mathbf{v}^{\perp} = \mathbf{v} - \hat{\mathbf{v}}$, 有正交分解 $\mathbf{v} = \hat{\mathbf{v}} + \mathbf{v}^{\perp}$, $\hat{\mathbf{v}} \perp \mathbf{v}^{\perp}$, 所以

$$\|\mathbf{v}\|^2 = \|\hat{\mathbf{v}}\|^2 + \|\mathbf{v}^{\perp}\|^2 \geq \|\hat{\mathbf{v}}\|^2 = \langle \mathbf{v}, \mathbf{u} \rangle^2 / \|\mathbf{u}\|^2.$$

■

命题 3.5 假设 V 是一个内积及向量空间，假设 $\mathbf{v} \in V$ 是任何一个向量。若 $\mathbf{u}_1, \ldots, \mathbf{u}_m \in V$ 相互正交，则 $P_U \mathbf{v} = \sum_{i=1}^{m} \left(\frac{\langle \mathbf{v}, \mathbf{u}_i \rangle}{\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \right) \mathbf{u}_i$. 是 $\mathbf{v}$ 在 $U = C(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ 上的投影。

证明：考虑 $\sum_{i=1}^{n} \lambda_i \mathbf{u}_i \in U$, 对任何 $k = 1, \ldots, m$, 令

$$0 = \langle \mathbf{v} - \sum_{i=1}^{n} \lambda_i \mathbf{u}_i, \mathbf{u}_k \rangle = \langle \mathbf{v}, \mathbf{u}_k \rangle - \sum_{i=1}^{m} \lambda_i \langle \mathbf{u}_i, \mathbf{u}_k \rangle = \langle \mathbf{v}, \mathbf{u}_k \rangle - \lambda_k \langle \mathbf{u}_k, \mathbf{u}_k \rangle,$$

因此 $\lambda_k = \langle \mathbf{v}, \mathbf{u}_k \rangle / \lambda_k \langle \mathbf{u}_k, \mathbf{u}_k \rangle$，对于这样选取的 λ's, $\sum_{i=1}^{n} \lambda_i \mathbf{u}_i$ 与 U 中任何向量正交，所以 $P_U \mathbf{v} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i$ 是 $\mathbf{v}$ 在 U 上的投影。
■

基于命题3.5，如果 $\mathbf{u}_1, \ldots, \mathbf{u}_m$ 是一组线性无关向量，则我们可通过 Gram-Schmidt 正交化变换得到相互正交的一组向量。

命题 3.6 （Gram-Schmidt 正交化）给定一组线性无关的向量 $\mathbf{u}_1, \ldots, \mathbf{u}_m \in V$, 下述 Gram-Schmidt 正交化变换得到相互正交的向量 $\mathbf{e}_1, \ldots, \mathbf{e}_m \in V$

$$\begin{aligned}
&\mathbf{e}_1 = \mathbf{u}_1 \\
&\mathbf{e}_2 = \mathbf{u}_2 - P_{\mathbf{e}_1}\mathbf{u}_2 = \mathbf{u}_2 - \left(\frac{\langle \mathbf{u}_2, \mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} \right) \mathbf{e}_1 \\
&\mathbf{e}_3 = \mathbf{u}_3 - P_{\mathbf{e}_1, \mathbf{e}_2}\mathbf{u}_3 = \mathbf{u}_3 - \left(\frac{\langle \mathbf{u}_3, \mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} \right) \mathbf{e}_1 - \left(\frac{\langle \mathbf{u}_3, \mathbf{e}_2 \rangle}{\langle \mathbf{e}_2, \mathbf{e}_2 \rangle} \right) \mathbf{e}_2 \\
&\cdots \\
&\mathbf{e}_m = \mathbf{u}_m - \left(\frac{\langle \mathbf{u}_m, \mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} \right) \mathbf{e}_1 - \cdots - \left(\frac{\langle \mathbf{u}_m, \mathbf{e}_{m-1} \rangle}{\langle \mathbf{e}_{m-1}, \mathbf{e}_{m-1} \rangle} \right) \mathbf{e}_{m-1}
\end{aligned}$$

Gram-Schmidt 正交化说明内积向量空间 V 的任何有限维子空间都存在一组正交基，进而命题3.3说明有限维子空间 U 存在

正交投影算子 P_U[2]。

命题 3.7 假设 U 是内积向量空间 V 的有限维子空间，则存在投影算子 P_U，使得对任何 $\mathbf{v} \in V$,

$$\mathbf{v} - P_U\mathbf{v} \perp U.$$

下一个例子说明实数域上实函数的三角级数逼近实际上是在三角函数空间上的投影。

例 3.4 (平方可积函数空间 L^2) 对任何 $f, g \in L^2$, 它们的内积

$$\langle f, g\rangle = \int f(x)g(x)dx.$$

考虑 $[0, 2\pi]$ 上的函数空间 $L^2([0, 2\pi])$。对任何 $f\ in V = L^2([0, 2\pi])$, 假设我们希望求解如下优化问题

$$\min_{a,b,c} ||f - f_{abc}||^2 = \min_{a,b,c} \int_0^{2\pi} \{f(x) - a - b(x) - c\sin(x)\}^2 dx$$

这是一个最小二乘问题,最优解为投影。假设函数 $1, \cos(\cdot), \sin(\cdot)$ 张成的向量空间 $V_0 = \{a + b\cos(\cdot) + c\sin(\cdot) : a, b, c \in R\}$,则任一 $f \in L^2([0, 2\pi])$ 在 V_0 上的投影 $\hat{f}(\cdot) = a + b\cos(\cdot) + c\sin(\cdot)$ 满足

$$\langle f - \hat{f}, 1\rangle = 0, \langle f - \hat{f}, \cos(\cdot)\rangle = 0, \langle f - \hat{f}, \sin(\cdot)\rangle = 0,$$

从上述三个方程分别解得

$$\hat{a} = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx,$$

$$\hat{b} = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(x)dx,$$

$$\hat{c} = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(x)dx,$$

进而得到投影 $\hat{f}(x) = \hat{a} + \hat{b}\cos(x)\hat{c}\sin(x)$。

扩大 V_0 的范围，我们可得到一般的 Fourier 三角级数逼近。

容易验证正交投影是 $V \to V$ 的线性变换，即 $P_U \in \mathcal{L}(V, V)$, 称为正交投影算子（映射、变换)。正交投影具有如下性质。

命题 3.8 假设 U 是向量空间 V 的向量子空间，假设存在 $V \to U$ 的投影映射 $P = P_U$。

(1) 如果正交投影存在，则必定唯一.

2: 更一般的结论是，如果 V 是希尔伯特空间, 若 U 是 V 的闭子空间, 则投影算子 P_U 必定存在。

(2) 对任何 $\mathbf{u} \in U$, $P\mathbf{u} = \mathbf{u}$.
(3) $P^2 = P$(幂等)，且 P 是自伴随的，即对任何 $\mathbf{v}, \mathbf{w} \in V$,

$$\langle P\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, P\mathbf{w} \rangle.$$

(4) 若算子 $T \neq 0 \in \mathcal{L}(V, V)$ 是自伴随、幂等的，则 T 一定是正交投影算子。

证明： (1) 若 $\hat{\mathbf{v}}_1 \in U$, $\hat{\mathbf{v}}_2 \in U$ 都是 $\mathbf{v}$ 在 U 空间中的投影, 则 $\mathbf{u} \triangleq \hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_2 \in U$，所以 $\langle \mathbf{v} - \hat{\mathbf{v}}_1, \mathbf{u} \rangle = 0$, $\langle \mathbf{v} - \hat{\mathbf{v}}_2, \mathbf{u} \rangle = 0$, 两式相减得到 $\langle \hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_2, \mathbf{u} \rangle = 0$, 此即 $\langle \hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_2, \hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_2 \rangle = 0$, 所以 $\hat{\mathbf{v}}_1 = \hat{\mathbf{v}}_2$.

(2) 若 $\mathbf{v} \in U$, 而它在 U 的投影 $P\mathbf{v}$ 也属于 U, 所以 $\mathbf{v} - P\mathbf{v} \in U$. 但根据正交投影的定义 $\mathbf{v} - P\mathbf{v} \in U$ 与 U 中的任何向量正交，从而与其自身正交，即 $\langle \mathbf{v} - P\mathbf{v}, \mathbf{v} - P\mathbf{v} \rangle = 0$, 所以 $P\mathbf{v} = \mathbf{v}$

(3) 对任何 $\mathbf{v} \in V$, $\hat{\mathbf{v}} = P\mathbf{v} \in U$, 由 (2), $P\hat{\mathbf{v}} = \hat{\mathbf{v}}$，这即是 $P(P\mathbf{v}) = P\mathbf{v}$, $\forall \mathbf{v} \in V$, 所以 $P^2 = P$。对任何 $\mathbf{v}, \mathbf{w} \in V$, 因为 $P\mathbf{w} \in U$, 所以

$$\langle \mathbf{v} - P\mathbf{v}, P\mathbf{w} \rangle = 0 \Leftrightarrow \langle \mathbf{v}, P\mathbf{w} \rangle = \langle P\mathbf{v}, P\mathbf{w} \rangle$$

同样因为 $P\mathbf{v} \in U$,

$$\langle \mathbf{w} - P\mathbf{w}, P\mathbf{v} \rangle = 0 \Leftrightarrow \langle \mathbf{w}, P\mathbf{v} \rangle = \langle P\mathbf{w}, P\mathbf{v} \rangle$$

所以

$$\langle P\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, P\mathbf{w} \rangle,$$

这说明 P 是自伴随的。

(4) 记 $U = TV$ 为 T 的像空间 (image,range), 则对任何 $\mathbf{u} \neq 0 \in U$, 假设 $\mathbf{u} = T\mathbf{w}, \mathbf{w} \in V$, 因为 $T^2 = T$, 我们有 $\mathbf{u} = T\mathbf{w} = T^2\mathbf{w} = T(T\mathbf{w}) = T\mathbf{u}$, 所以对任何 $\mathbf{v} \in V$,

$$\langle T\mathbf{v}, \mathbf{u} \rangle \overset{\text{自伴随}}{=} \langle \mathbf{v}, T\mathbf{u} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$$

这说明 $T\mathbf{v}$ 是 $\mathbf{v}$ 在子空间 U 上的投影。 ■

由上述结果，当 V 是有限为向量空间时，线性变换 P_U 对应的矩阵仍记作 P_U, 则该矩阵是对称幂等的 (对于矩阵，自伴随意味着对称)。我们将在后面详述 R^n 的投影。

3.4 随机变量空间

假设概率空间三件套 $(\Omega, \mathcal{F}, P)$，其中 Ω 是所有随机试验结果的集合，P 是概率测度，$\mathcal{F}$ 是 σ-域（可定义概率的 Ω 子集的集合)。假设 $x : \Omega \to R$ 是随机试验结果到实数域 R 的函数，如果对任何 $t \in R$,

$$\{\omega : x(\omega) \leq t\} \in \mathcal{F}.$$

我们称 x 是 $\mathcal{F}$-可测的，称为随机变量 [3] 。

按照定义3.3中函数空间的运算法则，对任何随机变量 x, y 定义加法 $x + y \in V$ 和数乘 $\lambda x \in V$, 则所有具有有限二阶矩的随机变量全体组成向量空间

$$V = L^2(\mathcal{F}) = \{x : x是\mathcal{F}\text{-可测随机变量}, E(x^2) < \infty\}.$$

对任何随机变量 $x, y \in V$, 定义内积

$$\langle x, y\rangle = E(xy)$$

(注：若 $P(x = 0) = 1$，则认为随机变量 $x = 0$)。

随机变量的子空间一般由特殊的随机变量，通常是具有较小可测范围的随机变量构成。假设 $\mathcal{G} \subset \mathcal{F}$ 是 $\mathcal{F}$ 的子 σ-域,

$$U = L^2(\mathcal{G}) = \{x : x是\mathcal{G}\text{-可测随机变量}, Ex^2 < \infty\} \subset V$$

是 V 的子空间。

定义 3.8 (条件期望) 任何 $y \in V$ 在 $U = L^2(\mathcal{G})$ 上的投影 $\hat{y}$ 满足

$$\langle y - \hat{y}, u\rangle = E(y - \hat{y})u = 0, \ \forall u \in U \tag{3.1}$$

即

$$E(yu) = E(\hat{y}u), \ \forall u \in U$$

该投影 $\hat{y}$ 应该记作 $P_U(y)$ 或 $P_{\mathcal{G}}(y)$, 但在概率论中该投影称为条件期望，习惯上记作 $\hat{y} = E(y|\mathcal{G})$（对于随机变量，后面将不再使用投影变换记号）。

定理 3.9 记 $\hat{y} = E(y|\mathcal{G})$，则

(1) $E(\hat{y}) = E(E(y|\mathcal{G})) = E(y)$.
(2) $\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(y - \hat{y}) = \mathrm{var}(E(y|\mathcal{G})) + E(\mathrm{var}(y|\mathcal{G}))$,
其中 $\mathrm{var}(y|\mathcal{G}) = E[(y - E(y|\mathcal{G}))^2|\mathcal{G}]$.

证明： (1) 等式 (3.1) 中取 $u = 1 \in U$, 有 $E(y - E(y|\mathcal{G})) = 0$.

(2) 等式 (3.1) 中取 $u = \hat{y} \in U$ 有 $E((y - \hat{y})\hat{y}) = 0$, 即 $\epsilon = y - \hat{y} \perp \hat{y}$, 所以有正交分解

$$y = \hat{y} + \epsilon$$

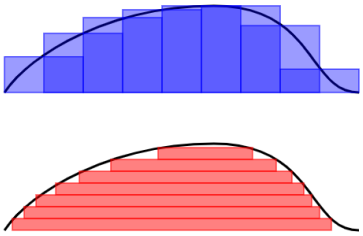
两边同时取方差，注意到 $\mathrm{var}(\epsilon) = E(var(y|\mathcal{G}))$. ■

一个重要的特殊情况是，对 $\mathbf{x} = (x_1, ..., x_p)^\top, x_i \in V$, 定义 $U = \{f(\mathbf{x}) : E(f(\mathbf{x})^2) < \infty\}$ 由所有 $\mathbf{x}$ 的（可测）函数构成。此时 $\mathcal{G} = \sigma(\mathbf{x})$, 投影写成 $\hat{y} = E(y|\mathbf{x})$。所以条件期望 $E(y|\mathbf{x})$ 可简

3: 随机变量作为随机结果的实函数，其定义域 Ω 上没有距离、相邻等概念, 但有概率大小的度量, 因而我们不能以无限划分定义域的方式定义积分, 但可以无限划分值域。例如勒贝格积分划分正函数 $x(\omega)$ 的值域，计算函数取值在 t 附近的横向长方形面积，即 $\{\omega : x(\omega) > t\}$ 的“长度”与 dt 的乘积，累加即为数学期望

$$E(x) = \sum P(\omega : x(\omega) > t)dt$$

其中要求集合 $\{\omega : x(\omega) > t\}$ 属于 $\mathcal{F}$, 可以定义“长度”即概率。



上图是黎曼积分，下图是勒贝格积分 (wiki).

单理解为 y 在所有 $\mathbf{x}$ 的函数构成的空间上的投影，解释为 y 中与 $\mathbf{x}$ 有关的部分，而

$$\Phi = \mathrm{var}(E(y|\mathbf{x}))/\mathrm{var}(y)$$

称为决定系数，是 $\mathbf{x}$ 所能解释的 y 的方差的比例。

例 3.5 （线性回归）定义3.8考虑了随机变量在特殊的可测函数空间上的投影，特别地 $E(y|\mathbf{x})$ 是随机变量 y 在随机向量 $\mathbf{x}$ 所有可测函数空间上的投影。在线性回归分析中，我们尤其关心响应变量 y 在自变量 $\mathbf{x}$ 的线性组合（而不是所有函数）空间上的投影。

(0) 假设 r.v.x 有二阶矩，即 $x \in V = L^2(\mathcal{F})$，记记常数随机变量 1 张成的空间 $U = \{a \times 1 : a \in R\} = R$, 则最小二乘问题

$$\min_{a\in R} \|x - a\|^2 = \min_{a\in R} E(x-a)^2$$

的最优解是 r.v. x 在实轴 R 上的投影 $\hat{a} = P_R x = E(x)$, 而 $x - E(x)$ 称为 x 的中心化，其长度平方 $E(x - E(x))^2 = var(x)$ 为方差。

(1) (简单线性回归) 假设 $x, y \in V = L^2(\mathcal{F})$, 考虑最小二乘问题

$$\min_{a,b\in R,} \|y - a - bx\|^2 = \min_{a,b\in R} E(y - a - bx)^2,$$

对参数 a, b 求导即可求出最优解，但我们从投影的观点求解如下:
$U = \{a + bx : a, b \in R\} \subset L^2(\mathcal{F})$ 为随机变量 x 和常数随机变量 1 张成的向量子空间，则随机变量 y 在 U 的投影 $\hat{y} = a + bx$ 满足（注意 $E(y - a - bx)^2$ 求导并令之为 0 得到同样的方程）

$$\langle y - a - bx, x\rangle = E(y - a - bx)x = 0,$$

$$\langle y - a - bx, 1\rangle = E(y - a - bx) = 0,$$

由此解得 $b = \Sigma_{yx}/\Sigma_{xx}, a = \mu_y - b\mu_x$, 其中 Σ_{yx}, Σ_{xx} 分别为 y, x 的协方差和 x 的方差，μ_z 为 z 的期望。所以

$$\hat{y} = \mu_y + \Sigma_{yx}/\Sigma_{xx}(x - \mu_x).$$

注：上述投影 $\hat{y} = P_x y$ 一般不等于条件期望 $E(y|x)$, 前者是在 $x, 1$ 线性组合张成的空间上的投影，后者是在所有 x 的可测函数组成的更大的空间上的投影。但当 (x, y) 服从二元正态时，上述投影 $\hat{y} = P_x y = E(y|x)$。

(2)（多变量线性回归）假设多个自变量 $\mathbf{x} = (x_1, ..., x_p)^\top, x_i \in V$, 假设二阶矩存在。考虑子空间

$$U = \{a + \mathbf{b}^\top \mathbf{x} : a \in R, \mathbf{b} \in R^p\} \subset V.$$

记 y 在 U 上的投影为 $\hat{y}$, 则 $y - \hat{y} = y - a - \mathbf{b}^\top \mathbf{x} \perp x_i, 1 \le i \le p$, 即 $\langle y - a - \mathbf{b}^\top \mathbf{x}, x_i \rangle = E(y - a - \mathbf{b}^\top \mathbf{x}) x_i = 0, 1 \le i \le p$, 这等价于

$$E\mathbf{x}(y - a - \mathbf{b}^\top \mathbf{x}) = 0$$

另外， $y - a - \mathbf{b}^\top \mathbf{x} \perp 1$, 即

$$\langle y - a - \mathbf{b}^\top \mathbf{x}, 1 \rangle = E(y - a - \mathbf{b}^\top \mathbf{x}) = 0,$$

由上述两个方程解得

$$\mathbf{b} = \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{x}y}, a = \mu_y - \mathbf{b}^\top \boldsymbol{\mu}_\mathbf{x}, \tag{3.2}$$

它们由 $y, \mathbf{x}$ 的均值和方差决定。所以

$$\hat{y} = \mu_y + \Sigma_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}).$$

注：我们称 $\epsilon \triangleq y - \hat{y} = y - \mu_y - \Sigma_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$, 或略去常数项 $y^\perp \triangleq y - \Sigma_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} \mathbf{x}$ 为 y 关于 $\mathbf{x}$ 的去相关化。称

$$y = \hat{y} + \epsilon = a + \mathbf{b}^\top \mathbf{x} + \epsilon$$

为线性回归模型， 其中 $a, \mathbf{b}$ 由 (3.2) 式给出， 而 $E(\epsilon) = 0, \text{var}(\epsilon) = \Sigma_{yy \bullet \mathbf{x}} \triangleq \Sigma_{yy} - \Sigma_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{x}y}$。

(3) (多元线性回归) 如果响应是随机向量 $\mathbf{y} = (y_1, ..., y_q)^\top$, 自变量依然是 $\mathbf{x}$, 那么我们可先求每个分量 y_i 的投影

$$\hat{y}_i = \mu_{y_i} + \Sigma_{y_i \mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}).$$

其中 $\Sigma_{y_i \mathbf{x}}$ 是 $1 \times p$ 行向量， 然后得到 $\mathbf{y}$ 的投影

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_q \end{pmatrix} = \begin{pmatrix} \mu_{y_1} \\ \vdots \\ \mu_{y_q} \end{pmatrix} + \begin{pmatrix} \Sigma_{y_1 \mathbf{x}} \\ \vdots \\ \Sigma_{y_q \mathbf{x}} \end{pmatrix} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$$

$$= \boldsymbol{\mu}_\mathbf{y} + \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$$

与 (2) 中的一元回归形式一致， 但这里 $\Sigma_{\mathbf{yx}} = \begin{pmatrix} \Sigma_{y_1 \mathbf{x}} \\ \vdots \\ \Sigma_{y_q \mathbf{x}} \end{pmatrix}$ 是一个 $q \times p$ 矩阵。

3.5 欧氏空间的投影矩阵

对任何 $\mathbf{x} = (x_i), \mathbf{y} = (y_i) \in R^n$，向量加法和数乘分别定义为

$$\mathbf{x} + \mathbf{y} = (x_i + y_i), \lambda\mathbf{x} = (\lambda x_i), \lambda \in R.$$

则 R^n 是一个向量空间。定义内积

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i,$$

模长 $||\mathbf{x}|| = (\mathbf{x}^\top \mathbf{x})^{1/2} = (\sum_{i=1}^{n} x_i^2)^{1/2}$。赋予内积的向量空间 R^n 称为欧氏空间。对于任何线性子空间 $V_0 \subset V = R^n$, 任何 $\mathbf{y} \in R^n$ 在 V_0 上的投影 $\hat{\mathbf{y}} = P_{V_0}\mathbf{y}$ 满足

$$\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{u} \rangle = (\mathbf{y} - \hat{\mathbf{y}})^\top \mathbf{u} = 0, \forall \mathbf{u} \in V_0,$$

投影是一个线性变换，在有限维空间 R^n 中，投影变换可以表示为投影矩阵。下面我们主要介绍 R^n 中的投影矩阵的求解及其性质。

任何 R^n 的子空间由一组线性向量张成，因此我们只需考虑矩阵的列向量张成的线性子空间。如果矩阵是不是列满秩的，即列向量不是线性无关的，则求解投影矩阵需要用到广义逆的概念，以及矩阵的奇异值分解。

3.5.1 奇异值分解

奇异值分解（SVD, Singular Value Decomposition）给出了一般矩阵行和列的特征刻画，是应用最为广泛的矩阵论结果。为了介绍 SVD，我们需要下述两个引理。

引理 3.10 (对称方阵的谱分解) 若 A 是一个 $n \times n$ 对称矩阵，则 A 的特征根全为实数，且所有特征向量可取为相互正交的模长为 1 的向量，即 A 有谱分解

$$A = V\Lambda V^\top,$$

其中 Λ 是 A 的所有特征根组成的对角矩阵，V 是正交矩阵 ($V^\top V = V^\top = I_n$), 其列向量为特征向量。

引理 3.11 对任何矩阵 A,

$$C(A) = C(AA^\top), C(A^\top) = C(A^\top A),$$

进而

$$rank(A) = rank(AA^\top) = rank(A^\top A) = rank(A^\top).$$

证明： 只需证 $C(A) = C(AA^\top)$. 显然 $C(AA^\top) \subset C(A)$, 假设 $\mathbf{x} \in C(AA^\top)^\perp$, 即 $AA^\top\mathbf{x} = 0$, 则 $\mathbf{x}^\top AA^\top\mathbf{x} = 0$ 即 $||A^\top\mathbf{x}||^2 = 0$, 所以 $A^\top\mathbf{x} = 0$, 这说明 $\mathbf{x} \in C(A)^\perp$, 所以 $C(AA^\top)^\perp \subset C(A)^\perp$, $C(A) \subset C(AA^\top)$. ■

假设 A 是一个方阵，其特征向量 $\mathbf{x}$ 满足特征方程

$$A\mathbf{x} = \lambda\mathbf{x}$$

λ 为对应的特征根，如果 λ 是非 0 实数，那么上述方程表明 $\mathbf{x} \in C(A)$, 所以特征向量代表了 A 的列空间 $C(A)$ 的某种特征。同样地，$A^\top$ 的特征向量刻画了 A 的行空间 $C(A^\top)$ 的特征。

对任何未必是方阵的矩阵 A，$AA^\top$ 是方阵，其非 0 特征根对应的特征向量刻画了 $C(AA^\top)$, 引理3.11表明这也是对 A 的列空间 $C(A)$ 的刻画；同样 $C(A^\top A)$ 的非 0 特征根对应的特征向量是对 A 的行空间 $C(A^\top)$ 的刻画。此外，两个方阵 $A^\top A$ 和 $AA^\top$ 有相同的非 0 特征根，这个共性将 $AA^\top$ 和 $A^\top A$ 的特征向量联系在一起就是 A 的奇异值分解。奇异值分解的存在性证明过程实际上就是将上述直观严格地表述出来。

定理 3.12 对任何秩为 r 的 $n \times m$ 矩阵 A，存在列正交矩阵 $U_{n\times r} = (\mathbf{u}_1, ..., \mathbf{u}_r)^\top$, $V_{m\times r} = (\mathbf{v}_1, ..., \mathbf{v}_r)^\top$ 和对角矩阵 $D_{r\times r} = \text{diag}(\sqrt{\lambda_1}..., \sqrt{\lambda_r})$, 满足 $U^\top U = V^\top V = I_r$, $\lambda_1 \geq ... \geq \lambda_r > 0$, 使得 A 有如下奇异值分解

$$A = UDV^\top = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top.$$

证明： 因为 $rank(A^\top A) = rank(A^\top) = r$, 所以 $A^\top A$ 是秩为 r 的 n 阶半正定矩阵, 假设其非 0 特征根为 $\lambda_1, ..., \lambda_r$, 对应的特征向量为 $\mathbf{v}_1, ..., \mathbf{v}_r$, 这些特征向量的模长取为 1 且可以取为是相互正交的（引理3.10），因此有特征方程:

$$A^\top A\mathbf{v}_i = \mathbf{v}_i\lambda_i,\ i = 1, ..., r,$$

两边同时左乘矩阵 A, 得 $AA^\top A\mathbf{v}_i = A\mathbf{v}_i\lambda_i$, 令 $\mathbf{u}_i = A\mathbf{v}_i/\sqrt{\lambda_i}$, 得

$$AA^\top\mathbf{u}_i = \mathbf{u}_i\lambda_i,\ i = 1, ..., r,$$

这说明 $\mathbf{u}_i$ 是 $AA^\top$ 对应于特征根 λ_i 的特征向量, 我们断言 $\mathbf{u}_1, ..., \mathbf{u}_1$ 模长为 1 且相互正交，这是因为 $\mathbf{u}_i^\top\mathbf{u}_i = \mathbf{v}_i^\top A^\top A\mathbf{v}_i/\lambda_i = \mathbf{v}_i^\top\mathbf{v}_i = 1$, 而对 $i \neq j$, $\mathbf{u}_i^\top\mathbf{u}_j = \mathbf{v}_i^\top A^\top A\mathbf{v}_j/\sqrt{\lambda_i\lambda_j} = \mathbf{v}_i^\top\mathbf{v}_j\sqrt{\lambda_j/\lambda_i} = 0$。记 $V = (\mathbf{v}_1, ..., \mathbf{v}_r)$, $U = (\mathbf{u}_1, ..., \mathbf{u}_r)$, 则 $V^\top V = U^\top U = I_r$, 且

$A\mathbf{v}_i = \mathbf{u}_i\sqrt{\lambda_i}, i = 1, ..., r$ 可以用矩阵表示为

$$AV = UD,$$

两边同时右乘 $V^\top$ 得 $AVV^\top = UDV^\top$, 下面只需证明 $AVV^\top = A$ (注意 $VV^\top$ 不是单位阵)。因为 $\mathbf{v}_1, ..., \mathbf{v}_r \in C(A^\top A)$, 且相互正交, 所以它们是 $C(A^\top A) = C(A^\top)$ 的一组标准正交基, 故存在 B 使得 $A^\top = VB$, 所以 $AVV^\top = B^\top V^\top VV^\top = B^\top V^\top = A$, 其中利用了 $V^\top V = I_r$。所以 $A = AVV^\top = UDV^\top$。 ■

定理3.12中的 U, V 的列都是相互正交且模长为 1 的矩阵, 将它们扩充为正交方阵即得到 SVD 的另一个版本。

推论 3.13 任一 $n \times m$ 矩阵 A 有奇异值分解: $A = U\Lambda V^\top$, 其中 U, V 分别是 n 阶、m 阶正交方阵, 即 $U^\top U = UU^\top = I_n$, $V^\top V = VV^\top = I_m$.

$$\Lambda_{n\times m} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

其中 $D = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ 是 $r \times r$ 对角阵, 其中 $r = rank(A)$, $\lambda_1 \geq ... \geq \lambda_r > 0$。

3.5.2 广义逆

考虑线性方程组 $A\mathbf{x} = \mathbf{b}$, 其中 A 是 $n \times m$ 矩阵, $\mathbf{x} \in R^m, \mathbf{b} \in R^n$, 方程组有解当且仅当 $\mathbf{b} \in C(A) = \{A\mathbf{t} : \mathbf{t} \in R^m\}$。若 $n = m$ 且 A 可逆, 则有唯一解 $\mathbf{x} = A^{-1}\mathbf{b}$; 若 A 不可逆甚至 A 不是方阵, 如果方程有解, 这些解能否类似地表示为 $\mathbf{x} = A^{inv}\mathbf{b}$? 其中 A^{inv} 是 A 的一类 "逆"。

假设存在某个 "逆" 矩阵 A^{inv}, 使得 $\mathbf{x} = A^{inv}\mathbf{b}$ 是方程的解, 则 $\mathbf{b} = A\mathbf{x} = A(A^{inv}\mathbf{b}) = A(A^{inv}(A\mathbf{t}))$。因为方程有解, 则 $\mathbf{b} \in C(A)$, 必存在 $\mathbf{t} \in R^m$ 使得 $\mathbf{b} = A\mathbf{t}$。因此我们有 $\mathbf{b} = A\mathbf{t} = AA^{inv}A\mathbf{t}$。所以我们可以断言, 如果 A^{inv} 满足 $AA^{inv}A = A$, 则 $\mathbf{x} = A^{inv}\mathbf{b}$ 是方程的解。基于上述考虑, 定义广义逆如下。

定义 3.9 假设 A 是 $n \times m$ 矩阵, 任何满足

$$AXA = A$$

的 $m \times n$ 矩阵 X 称为 A 的广义逆。

广义逆一定存在, 但不唯一。下述命题利用奇异值分解给出了广义逆的一般表达形式。

命题 3.14 假设 A 的奇异值分解为 $A = U\Lambda V^\top$, 其中 U,V 分别是 n 阶、m 阶正交方阵，

$$\Lambda_{n\times m} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

其中 $D > 0$ 是 $r \times r$ 对角阵，$r = rank(A)$。则任何广义逆具有如下形式

$$A^- = V\begin{pmatrix} D^{-1} & * \\ * & * \end{pmatrix}U^\top$$

其中 $*$ 处可以是任意矩阵。

证明： 若 X 是 A 的任一广义逆，则 $AXA = A$, 即

$$U\Lambda V^\top XU\Lambda V^\top = U\Lambda V^\top$$

令 $Y = V^\top XU = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix}$, 其中 Y_{11} 是 $r \times r$ 矩阵。则上式市等价于

$$\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix}\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

由此知 $Y_{11} = D^{-1}$, 其它子块任意。 ■

定理 3.15 若方程 $A\mathbf{x} = \mathbf{b}$ 有解（$\mathbf{b} \neq 0$），则所有解都具有形式 $\mathbf{x} = A^-\mathbf{b}$.

证明： 任取一个 A 的广义逆矩阵 B, 则 $\mathbf{x}_0 = B\mathbf{b}$ 是一个解，即 $A\mathbf{x}_0 = AB\mathbf{b} = \mathbf{b}$。假设 $\mathbf{x}$ 是方程的任一解, 则 $A\mathbf{x} - A\mathbf{x}_0 = \mathbf{b} - \mathbf{b} = 0$, 记 $\mathbf{v} = \mathbf{x} - \mathbf{x}_0$, 则 $A\mathbf{v} = 0$, 且

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{v} = B\mathbf{b} + \mathbf{v}\frac{\mathbf{b}^\top\mathbf{b}}{\mathbf{b}^\top\mathbf{b}} = \left(B + \frac{\mathbf{v}\mathbf{b}^\top}{\mathbf{b}^\top\mathbf{b}}\right)\mathbf{b} \triangleq C\mathbf{b}$$

其中 $C = B + \frac{\mathbf{v}\mathbf{b}^\top}{\mathbf{b}^\top\mathbf{b}}$, 而 $ACA = ABA = A\mathbf{v}\mathbf{b}^\top A/\|\mathbf{b}\|^2 + ABA = A$, 所以 C 是 A 的广义逆，从而 $\mathbf{x}$ 具有 $A^-\mathbf{b}$ 的形式。 ■

3.5.3 正交投影矩阵

定理 3.16 假设 A 是 $n \times m$ 矩阵，A 的列向量张成的空间

$$V_0 = C(A) = \{A\mathbf{t} : \mathbf{t} \in R^m\}.$$

设 $\mathbf{y} \in R^n$, 它在 V_0 上的投影为

$$\hat{\mathbf{y}} = P_A\mathbf{y} = A(A^\top A)^- A^\top \mathbf{y},$$

其中 $P_A = A(A^\top A)^- A^\top$ 是投影线性变换对应的投影矩阵，称为 A 对应的投影矩阵。

证明： 因为 $\mathbf{y}$ 在 V_0 上的投影 $\hat{\mathbf{y}} \in V_0$, 故存在 $\mathbf{t} \in R^m$, 使得 $\hat{\mathbf{y}} = A\mathbf{t}$。又因为

$$\mathbf{y} - \hat{\mathbf{y}} \perp V_0$$

故对任何 $A\mathbf{s} \in V_0$ (任何 $\mathbf{s} \in R^m$) 有 $\mathbf{y} - \hat{\mathbf{y}} \perp A\mathbf{s}$, 即

$$0 = \langle \mathbf{y} - \hat{\mathbf{y}}, A\mathbf{s} \rangle = \langle A^\top(\mathbf{y} - \hat{\mathbf{y}}), \mathbf{s} \rangle$$

这说明 $A^\top(\mathbf{y} - \hat{\mathbf{y}}) \in R^m$ 与任何 $\mathbf{s} \in R^m$ 正交，所以必定[4]

$$A^\top(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$$

即

$$A^\top\mathbf{y} = A^\top\hat{\mathbf{y}} = A^\top A\mathbf{t}$$

因为 $A^\top\mathbf{y} \in C(A^\top) = C(A^\top A)$, 所以方程必定有解

$$\mathbf{t} = (A^\top A)^- A^\top \mathbf{y}$$

所以投影为

$$\hat{\mathbf{y}} = A(A^\top A)^- A^\top \mathbf{y} \triangleq P_A\mathbf{y}.$$

■

注：需要注意的是，若 $A^\top A$ 不可逆, 则解 $\mathbf{t}$ 不唯一，但根据投影的唯一性，$P_A = A(A^\top A)^- A^\top$ 是唯一的，即 P_A 不依赖于广义逆 $(A^\top A)^-$ 的具体选择。

命题3.8已经说明了内积向量空间的投影是唯一的和对称幂等的，而且具有最小二乘特性。作为特殊情况，欧氏空间的投影当然也具备这些性质。下面使用矩阵理论再次给予证明。

命题 3.17 假设 A 是 $n \times m$ 矩阵，投影矩阵 $P_A = A(A^\top A)^- A^\top$, 则

(1) P_A 唯一，与广义逆的选择无关。
(2) P_A 是对称幂等矩阵。反之，任一对称幂等矩阵是投影阵。
(3) $I_n - P_A$ 是子空间 $C(A)^\perp$ 的投影矩阵，且 $P_A(I_n - P_A) = 0$。
(4) 按列划分 $A = (A_1, A_2)$, 若 $A_1^\perp A_2 = 0$, 则 $P_A = P_{A_1} + P_{A_2}$。
(5) 对任何 $\mathbf{y} \in R^n$,

$$P_A\mathbf{y} = arg \min_{\mathbf{u} \in C(A)} \|\mathbf{y} - \mathbf{u}\|^2.$$

4: $A^\top$ 是 A 的伴随 (adjoint) 矩阵:

$$\langle A\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A^\top\mathbf{v} \rangle.$$

证明： (1) 因为 $C(A^\top) \subset C(A^\top A)$, 故存在矩阵 B 使得 $A^\top = (A^\top A)B$, 所以 $P_A = A(A^\top A)^- A^\top = B^\top A^\top A(A^\top A)^- A^\top AB = B^\top A^\top AB$，不依赖于广义逆的具体选择。

(2) $P_A^2 = A(A^\top A)^- A^\top A(A^\top A)^- A^\top$ 中将最右端的 $A^\top$ 替换为 $AA^\top B$ 即可得到 $P_A^2 = P_A$。反之，若 C 是任一 $n \times n$ 对称幂等矩阵，$rank(C) = r$, 则其所有特征根为 r 个 1 和 $n - r$ 个 0，由对称矩阵的谱分解知存在 n 阶正交矩阵 O 使得

$$C = O \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} O^\top$$

划分 $O = (O_1, O_2)$, 其中 O_1，O_2 的列数分别是 r 和 $n - r$, 由上述谱分解得 $C = (O_1, O_2) \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} O_1^\top \\ O_2^\top \end{pmatrix} = O_1 O_1^\top$。另一方面，由 $O^\top O = I_n$ 知 $O_1^\top O_1 = I_r$ 所以 $C = O_1 O_1^\top = O_1 (O_1^\top O_1)^{-1} O_1^\top$ 具有投影矩阵的形式，是一个投影阵。

(3) 任何 $\mathbf{u} \in C(A)^\perp$ 与 $C(A)$ 正交，特别地对任何任何 $\mathbf{y} \in R^n$

$$\langle \mathbf{u}, P_A \mathbf{y} \rangle = \langle \mathbf{u}, \mathbf{y} - (I_n - P_A)\mathbf{y} \rangle = 0$$

所以 $I_n - P_A$ 是 $C(A)^\perp$ 对应的投影矩阵。显然 $(I_n - P_A)_A P = P_A - P_A^2 = 0$。

(4) $A^\top A = \begin{pmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{pmatrix}$，因为投影阵与广义逆的具体选择无关，我们取 $(A^\top A)^- = \begin{pmatrix} (A_1^\top A_1)^- & 0 \\ 0 & (A_2^\top A_2)^- \end{pmatrix}$, 则容易验证这确实是一个广义逆，从而

$$P_A = (A_1, A_2) \begin{pmatrix} (A_1^\top A_1)^- & 0 \\ 0 & (A_2^\top A_2)^- \end{pmatrix} \begin{pmatrix} A_1^\top \\ A_2^\top \end{pmatrix} = P_{A_1} + P_{A_2}.$$

(5) 对任何 $\mathbf{y} \in R^n$, $\mathbf{u} \in C(A)$, $\mathbf{y} - P_A \mathbf{y} \perp C(A)$, 而 $P_A \mathbf{y} - \mathbf{u} \in C(A)$, 所以

$$\begin{aligned} \|\mathbf{y} - \mathbf{u}\|^2 &= \|(\mathbf{y} - P_A \mathbf{y}) + (P_A \mathbf{y} - \mathbf{u})\|^2 \\ &= \|\mathbf{y} - P_A \mathbf{y}\|^2 + \|P_A \mathbf{y} - \mathbf{u}\|^2 \geq \|\mathbf{y} - P_A \mathbf{y}\|^2. \end{aligned}$$

当 $\mathbf{u} = P_A \mathbf{y}$ 时达到最小。 ■

鸣谢

感谢卞泽宇同学阅读本章的初稿并提出若干宝贵意见。恳请各位指出其中存在的错误、进一步修改的建议、需要删减或增加的内容。

参考文献

Here are the references in citation order.

[1] S. Axler. *Linear Algebra Done Right*. Springer, 2015 (cited on page 4).

按字母排序的索引

Cauchy-Schwarz 不等式, 9

Gram-Schmidt 正交化, 9

SVD, 16, 17

内积向量空间, 6, 8
函数空间, 4
函数空间 L^2, 7, 10
向量空间, 3
向量空间：函数空间, 4
奇异值分解, 16, 17
广义逆, 17

投影 (条件期望), 12

最小二乘, 8, 19
有限维向量空间, 5
条件期望, 12
欧氏空间, 7, 15
正交投影 (一般内积向量空间), 8
正交投影矩阵, 18, 19
毕达哥拉斯定理, 6

线性变换, 4
线性回归 (随机变量投影), 13

谱分解, 15

随机变量空间, 7, 11